PROYECTO DE CALENTAMIENTO GLOBAL

Nahuel Acevedo y
Maria Carbajal

Comisión 29800

Profesor Enrique Escalante Notario

CODER HOUSE

Contenido

Contenido

i) Exploración de Dataset

iii) Preguntas//hipótesis

iv) EDA - Exploratory Data Analysis
- Visualización de Datos con Seaborn (sns)
- Análisis multivariado
- Pandas Profiling

v) Recomendaciones con base en Insights observados
- Regresion lineal - para predicción de variables numéricas continuas

v) Preprocesamiento de Datos

vi) Splitting de dataset

vii) Modelado de Machine Learning
- Algoritmo Random Forest
- Algoritmo SVR
- Algoritmo K-Means

viii) Explicación de Modelos
- Visualización de Resultados de Modelo

ix) Insights y recomendaciones

x) Fuentes

Abstracto

El presente estudio trata de responder a los siguientes interrogantes: ¿Estamos ante una crisis climática, esto es, el aumento sostenido de la variación de la temperatura? ¿Es una tendencia que se mantendrá a lo largo de los próximos años? ¿Como será su evolución en Argentina y en la región en los próximos años? La principal hipótesis que tratamos de comprobar es que, actualmente, estamos viviendo una crisis climática tal como sugieren distintos estudios científicos realizados por la ONU. Para responder este interrogante se llevó a cabo el análisis estadístico de los datos proporcionados por GISTEMP y NASA-GISS. A partir de este, obtuvimos un gráfico de serie de tiempo que nos permite observar el patrón alcista de la temperatura acumulada en el rango 1961-2019.

Objetivo

Motivados por las recientes temperaturas que se están presentando en el mundo que son sin precedentes queremos ofrecer un estudio que ayude a empresas e instituciones gubernamentales a prepararse para afrontar esta situación. Nuestro objetivo es por medio de regresión estimar futuras temperaturas de modo tal que sea predecible y poder actuar en modo proactivo.

Contexto comercial

En la sección anterior se puede apreciar como este cambio en la temperatura ha forzado a instituciones y gobiernos a implementar medidas de forma repentina por lo que además de conocer un valor futuro para la temperatura es muy util conocer la correlación que pueden tener con otros fenómenos como:

- Deforestación
- Emisiones de C02
- Desertificación
- Energía usada

Problema comercial

Se buscaría prever los cambios o impactos de estas temperaturas en caso salgan de lo normal. Ejemplo, actualmente en Inglaterra donde no se tienen jamás 40 grados han tenido que envolver puentes y pintar vías de tren en blanco para evitar la dilatación de los materiales. En Europa en general a estas temperaturas se refuerza la atención a a adultos

mayores que sufren de problemas de deshidratación en estas épocas. Por lo que de cara a lo que podamos predecir para este verano se podría tomar nota de estas acciones.

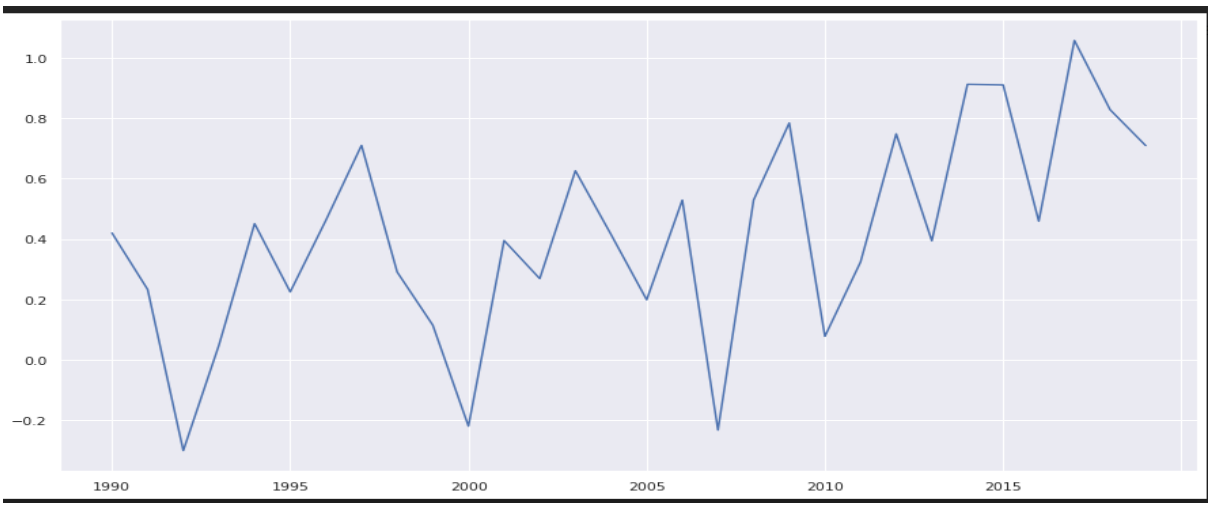
Contexto analítico

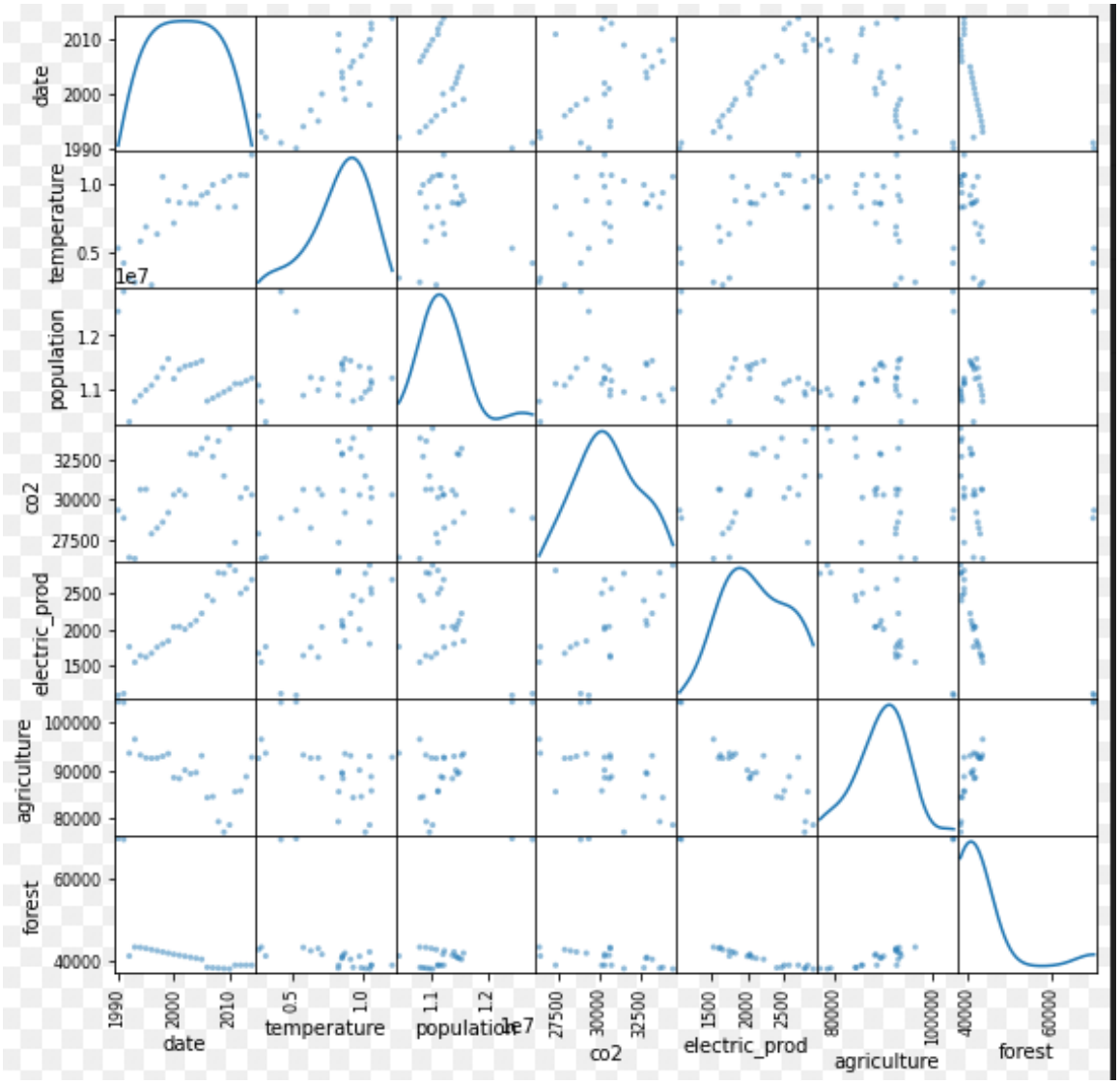
Buscaríamos más datasets para buscar correlaciones en: https://www.worldometers.info/

El actual dataset que se trabajó cubre el período 1961-2019. Las estadísticas están disponibles para las anomalías de temperatura media mensual, estacional y anual, es decir, el cambio de temperatura con respecto a una climatología de referencia, correspondiente al período 1951-1980. Los datos se basan en los datos GISTEMP disponibles públicamente, los datos de cambio de temperatura de la superficie global distribuidos por el Instituto Goddard de Estudios Espaciales de la Administración Nacional de Aeronáutica y del Espacio (NASA-GISS).

i) Exploración de Dataset

Encontramos sólidez en los datos a partir de 1990 por lo que elegimos este periodo para hacer el estudio. A partir de 2007 se ve una nueva tendencia de alza y cambio positivo de temperatura.





Hicimos un análisis multivariado para contrastar el efecto de las variables de interés entre sí.

Vemos tendencias bastante obvias **como el aumento de producción eléctrica yendo de la mano con el aumento de población**, el uso de la variable fecha es bastante coherente pues muestra la evolución de las variables en el tiempo.

Variables como el aumento en emisiones de CO2 y la producción de energía eléctrica presentan una correlación positiva con el aumento de temperatura.

ii) Data Wrangling

El dataset contiene las siguientes variables:

-Access to electricity (% of population).
-Agricultural land (% of land area).
-Annual freshwater withdrawals, total (billion cubic meters).
-CO2 emissions (kg per 2015 US$ of GDP).
-CO2 emissions from liquid fuel consumption (kt) .
-Droughts, floods, extreme temperatures (% of population, average 1990-2009).

-Ease of doing business rank (1=most business-friendly regulations).
-Electric power consumption (kWh per capita).
-Electricity production from coal sources (% of total).
-Electricity production from hydroelectric sources (% of total).
-Electricity production from natural gas sources (% of total).
-Electricity production from nuclear sources (% of total).
-Energy use (kg of oil equivalent) per $1,000 GDP (constant 2017 PPP).
-Forest area (sq. km) .
-HFC gas emissions (thousand metric tons of CO2 equivalent).
-Land area where elevation is below 5 meters (% of total land area).
-Methane emissions (% change from 1990).
-Methane emissions (kt of CO2 equivalent).
-Mortality rate, under-5 (per 1,000 live births).
-Nitrous oxide emissions (% change from 1990).
-Nitrous oxide emissions (thousand metric tons of CO2 equivalent).
-Other greenhouse gas emissions (% change from 1990).
-Population growth (annual %).
-Rural population living in areas where elevation is below 5 meters (% of total population).
-Temperature change.
-Terrestrial and marine protected areas (% of total territorial area).
-Total greenhouse gas emissions (kt of CO2 equivalent).
-Urban land area where elevation is below 5 meters (% of total land area).

Por otra parte, tenemos en cuenta que nuestro dataset es mundial y decidimos explorar Argentina realizando operaciones para obtener un dataframe con las variables que mencionamos anteriormente.

Realizamos una limpieza en el dataset con el fin de no obtener valores nulos:

```
df_agri['date'] = df_agri['date'].astype(float)
df_fore['date'] = df_fore['date'].astype(float)
df_elec['date'] = df_elec['date'].astype(float)
df_co2e['date'] = df_co2e['date'].astype(float)
df_popu['date'] = df_popu['date'].astype(float)

df_popu.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11704 entries, 0 to 11703
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Area        11704 non-null  object
 1   name        11704 non-null  object
 2   date        11704 non-null  float64
 3   population  11635 non-null  float64
dtypes: float64(2), object(2)
memory usage: 365.9+ KB
```

iii) Preguntas//hipótesis

El presente estudio trata de responder a los siguientes interrogantes: ¿Qué factores afectaron la temperatura en Argentina? ¿Qué factores impactan más en el efecto invernadero?¿Cómo incide el CO2 en el calentamiento global? ¿Qué actividad industrial/económica que haya emergido en el tiempo puede relacionarse con el efecto invernadero?

La principal hipótesis que tratamos de comprobar es que, actualmente, estamos viviendo una crisis climática tal como sugieren distintos estudios científicos realizados por la ONU. Para responder este interrogante se llevó a cabo el análisis estadístico de los datos proporcionados por GISTEMP y NASA-GISS. A partir de este, obtuvimos un gráfico de serie de tiempo que nos permite observar el patrón alcista de la temperatura acumulada en el rango 1961-2019.

Objetivo

Motivados por las recientes temperaturas que se están presentando en el mundo que son sin precedentes queremos ofrecer un estudio que ayude a empresas e instituciones gubernamentales a prepararse para afrontar esta situación. Nuestro objetivo es por medio de regresión estimar futuras temperaturas de modo tal que sea predecible y poder actuar en modo proactivo.

Contexto comercial

En la sección anterior se puede apreciar como este cambio en la temperatura ha forzado a instituciones y gobiernos a implementar medidas de forma repentina por lo que además de conocer un valor futuro para la temperatura es muy util conocer la correlación que pueden tener con otros fenómenos como:

Deforestación Emisiones de C02 Desertificación Energía usada Problema comercial

Se buscaría prever los cambios o impactos de estas temperaturas en caso salgan de lo normal. Ejemplo, actualmente en Inglaterra donde no se tienen jamás 40 grados han tenido que envolver puentes y pintar vías de tren en blanco para evitar la dilatación de los materiales. En Europa en general a estas temperaturas se refuerza la atención a a adultos mayores que sufren de problemas de deshidratación en estas épocas. Por lo que de cara a lo que podamos predecir para este verano se podría tomar nota de estas acciones.

Contexto analítico

Buscaríamos más datasets para buscar correlaciones en: https://www.worldometers.info/

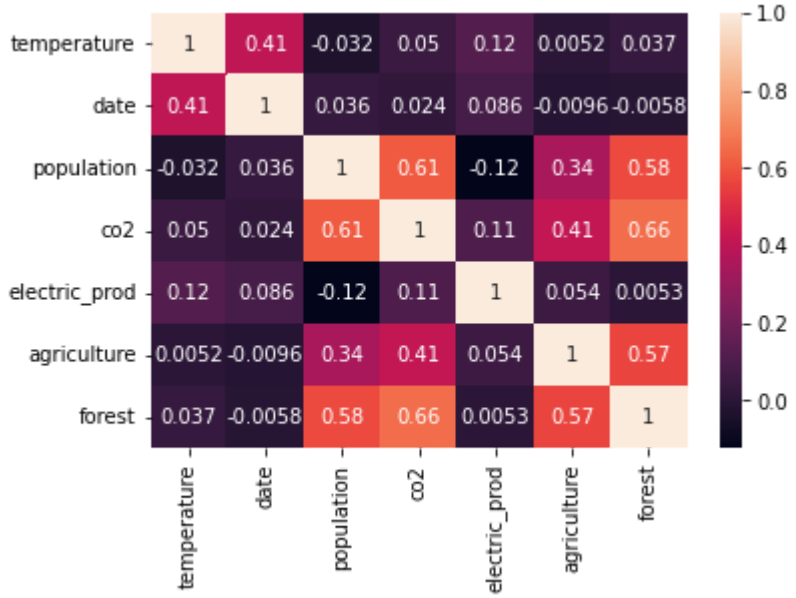
El actual dataset que se trabajó cubre el período 1961-2019. Las estadísticas están disponibles para las anomalías de temperatura media mensual, estacional y anual, es decir, el cambio de temperatura con respecto a una climatología de referencia, correspondiente al

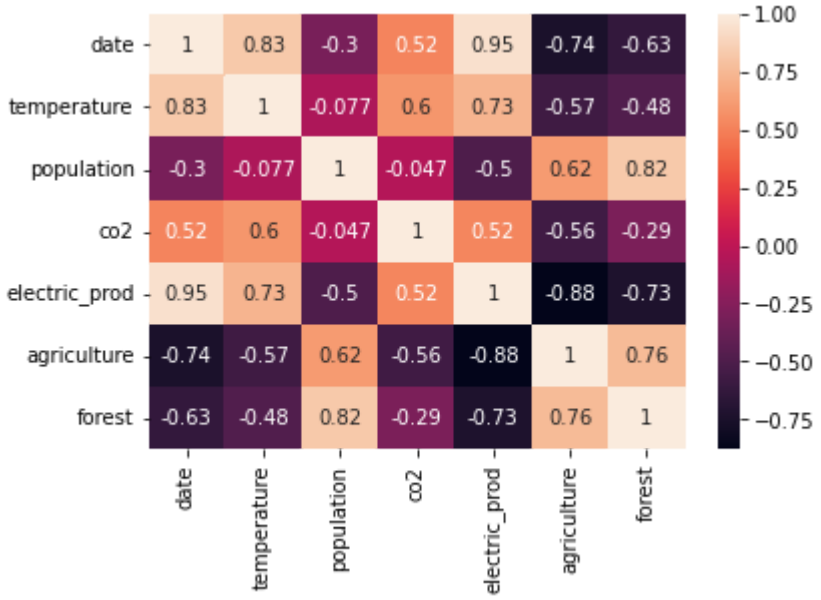
período 1951-1980. Los datos se basan en los datos GISTEMP disponibles públicamente, los datos de cambio de temperatura de la superficie global distribuidos por el Instituto Goddard de Estudios Espaciales de la Administración Nacional de Aeronáutica y del Espacio (NASA-GISS).

A continuación vamos a analizar datos usando distintas agrupaciones para darnos una idea de cómo están distribuidos los datos
Hipótesis: El aumento del efecto invernadero está relacionado con el aumento del uso de electricidad y de la emisión de co2.

iv) EDA - Exploratory Data Analysis

Visualización de Datos con Seaborn (sns)





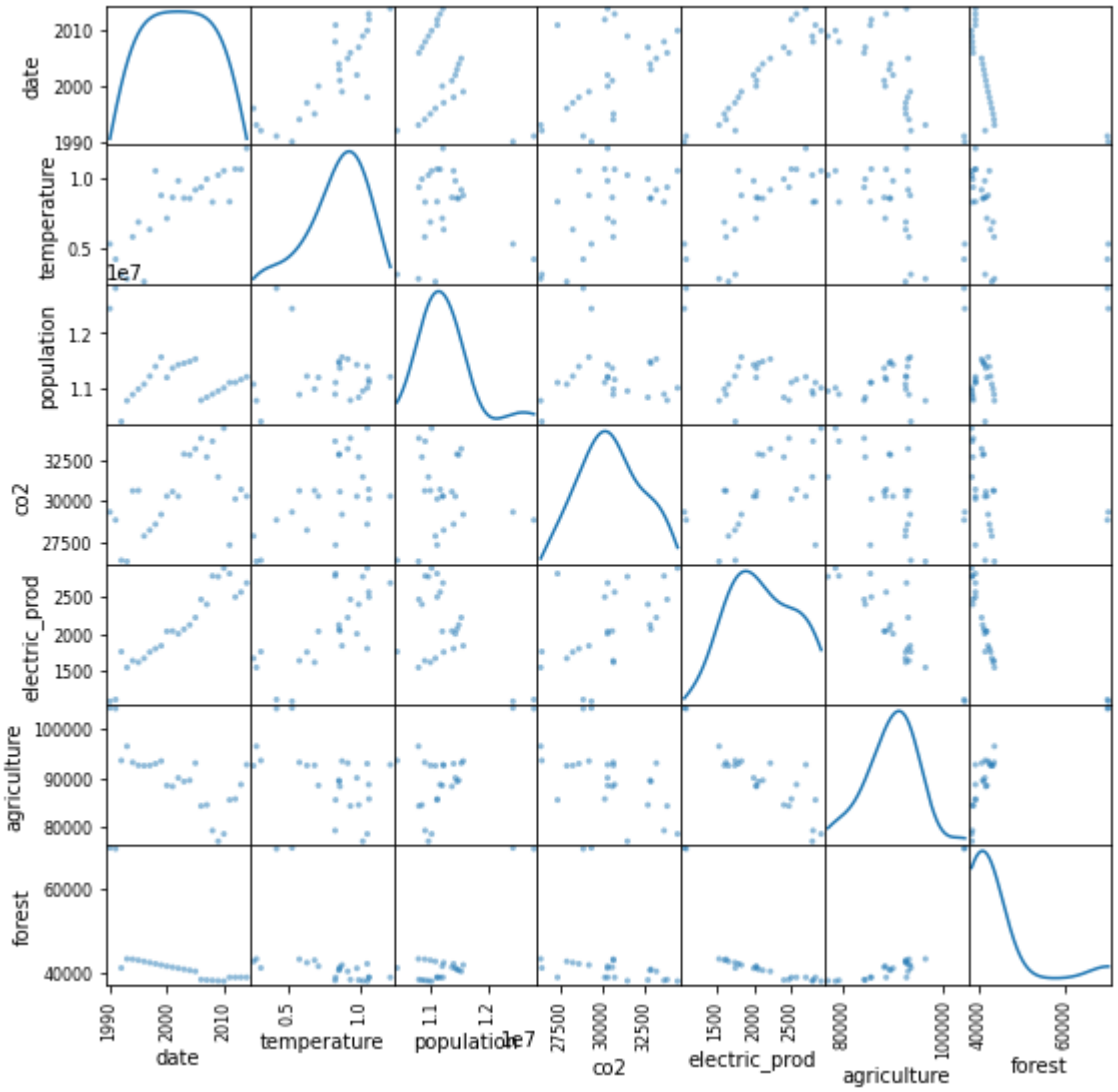
Análisis multivariado

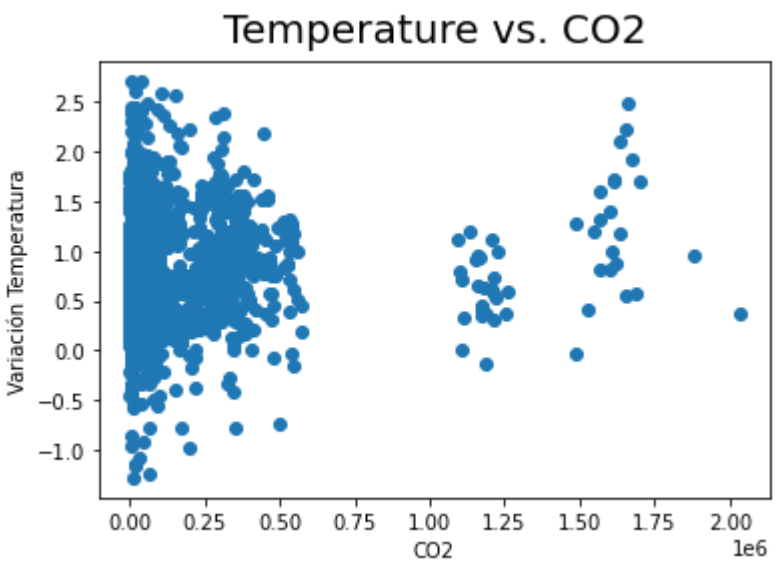
Hicimos un análisis multivariado para contrastar el efecto de las variables de interés entre sí.

Vemos tendencias bastante obvias como el aumento de producción eléctrica yendo de la mano con el aumento de población, el uso de la variable fecha es bastante coherente pues muestra la evolución de las variables en el tiempo.

¿Cómo incide el CO2 en la temperatura?

Variables como el aumento en emisiones de CO2 y la producción de energía eléctrica presentan una correlación positiva con el aumento de temperatura.


date
temperature
population
co2
electric_prod
agriculture
forest
1e7


Temperature vs. CO2
Variación Temperatura
CO2
1e6

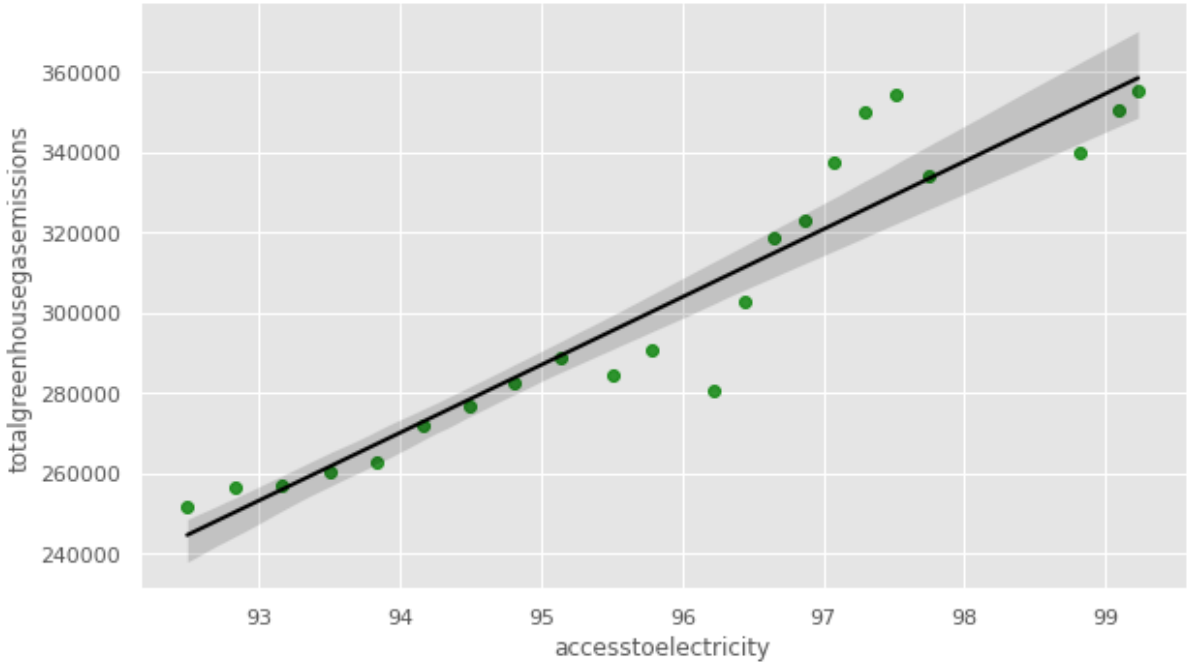
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            temperature   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     3.731
Date:                Wed, 11 Jan 2023   Prob (F-statistic):             0.0536
Time:                        23:22:49   Log-Likelihood:                -1306.2
No. Observations:                1521   AIC:                             2616.
Df Residuals:                    1519   BIC:                             2627.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.8322      0.016     50.763      0.000       0.800       0.864
co2         1.067e-07   5.52e-08      1.932      0.054   -1.65e-09    2.15e-07
==============================================================================
Omnibus:                       18.536   Durbin-Watson:                   1.512
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.121
Skew:                           0.096   Prob(JB):                     7.83e-07
Kurtosis:                       3.638   Cond. No.                     3.32e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.32e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Lo que podemos observar en la matriz es que tanto, el aumento de los años, el crecimiento de la población y la producción de electricidad tienen correlación positiva con el aumento de temperatura así como los kilómetros cuadrados de bosque y la agricultura con inversamente proporcionales (correlación negativa).

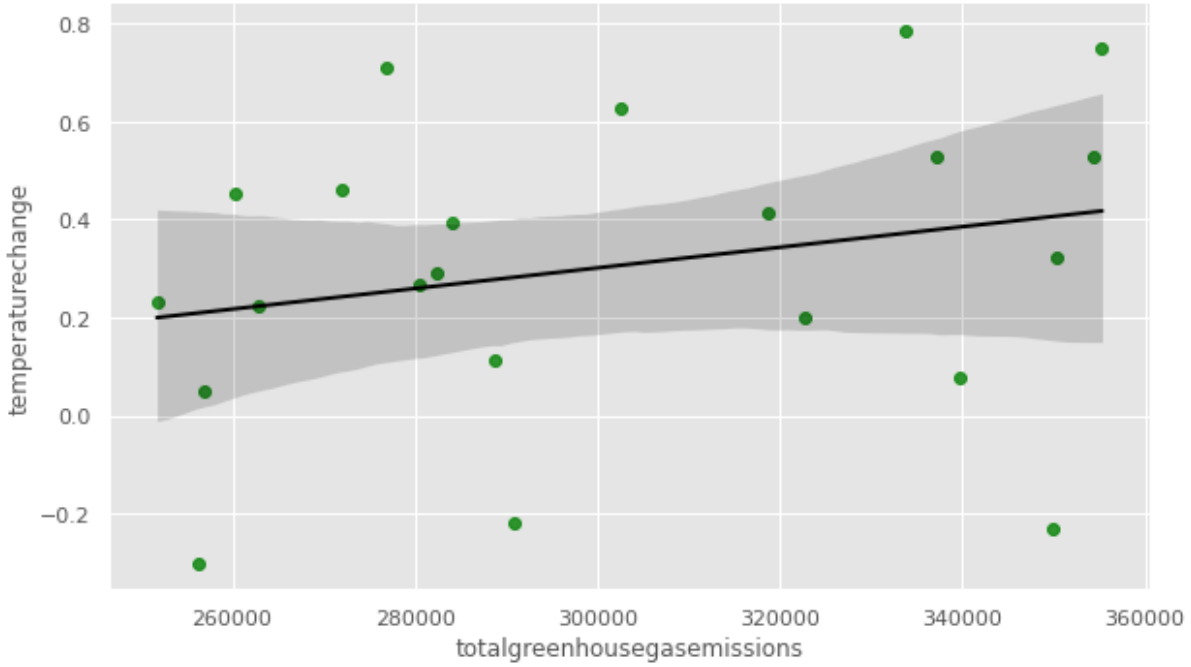
Pandas Profiling

El reporte de profiling ha sido muy útil y revelador tanto que queremos explorar mejor la variable de cambio de temperatura ya que al parecer no se relaciona directamente con las dependientes. En la sección interactions vemos que la mejor es la de emisión de gases de efecto invernadero.

En el siguiente gráfico, podemos ver que el acceso a la electricidad es la variable más robusta:



En la siguiente exploración observamos que el cambio de temperatura es muy disperso y no nos permite seguir un patrón.



v) Recomendaciones con base en Insights observados

Regresion lineal - para predicción de variables numéricas continuas

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     totalgreenhousegasemissions   R-squared:                       0.898
Model:                                     OLS   Adj. R-squared:                  0.893
Method:                          Least Squares   F-statistic:                     176.5
Date:                         Wed, 11 Jan 2023   Prob (F-statistic):           2.20e-11
Time:                                 23:23:06   Log-Likelihood:                -236.50
No. Observations:                           22   AIC:                             477.0
Df Residuals:                               20   BIC:                             479.2
Df Model:                                    1
Covariance Type:                     nonrobust
==========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept           -1.319e+06   1.22e+05    -10.813      0.000   -1.57e+06   -1.06e+06
accesstoelectricity   1.69e+04   1272.171     13.286      0.000    1.42e+04    1.96e+04
==============================================================================
Omnibus:                        2.498   Durbin-Watson:                   0.671
Prob(Omnibus):                  0.287   Jarque-Bera (JB):                0.980
Skew:                           0.279   Prob(JB):                        0.613
Kurtosis:                       3.871   Cond. No.                     4.64e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.64e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

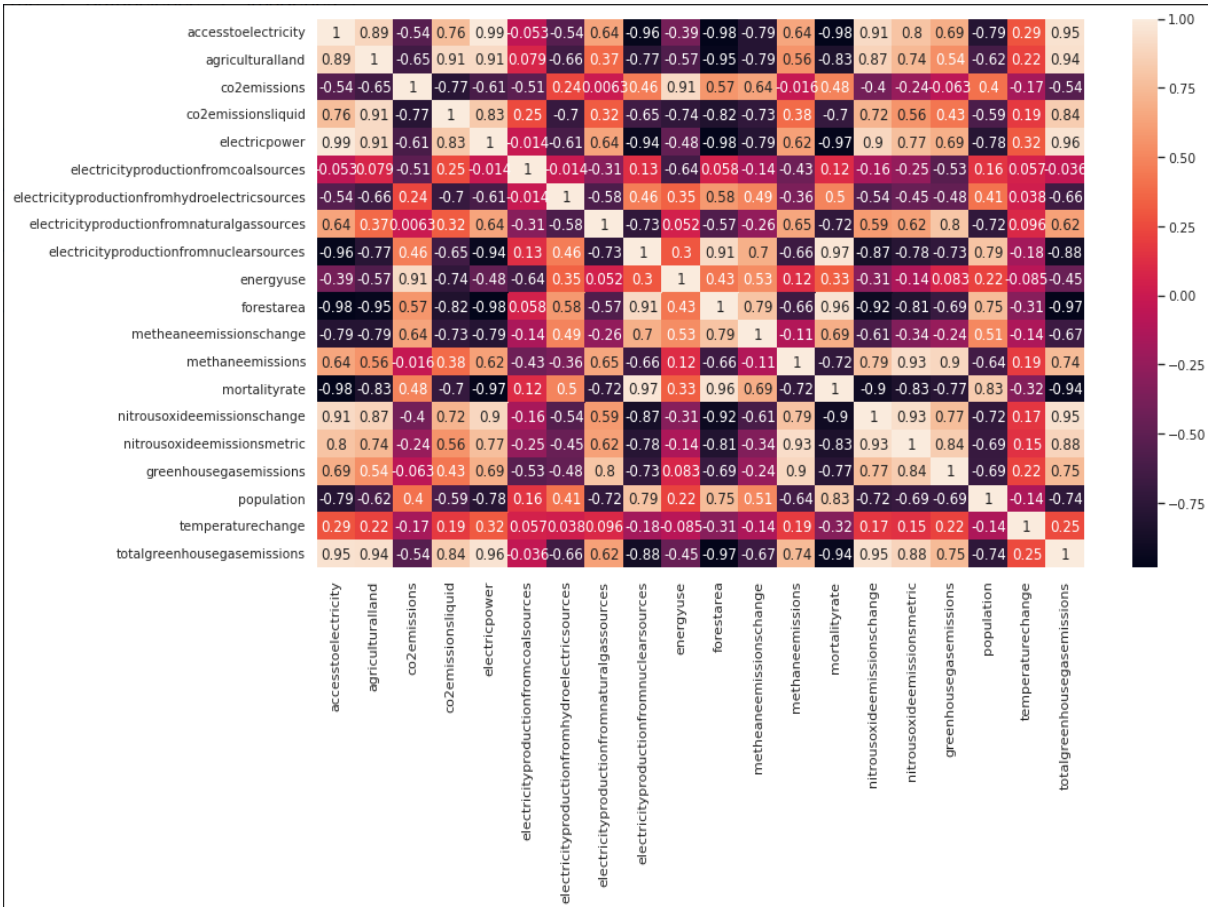
El resultado de R^2 es de 0.89 por lo que podemos interpretar que el 89% de los datos de acceso a la electricidad puede explicar elrresultado de las emisiones de gas invernadero, siendo esta una variable sólida.

Regresion Lineal Multiple

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     totalgreenhousegasemissions   R-squared:                       0.970
Model:                                     OLS   Adj. R-squared:                  0.961
Method:                          Least Squares   F-statistic:                     103.8
Date:                         Wed, 11 Jan 2023   Prob (F-statistic):           1.30e-11
Time:                                 23:23:06   Log-Likelihood:                -223.02
No. Observations:                           22   AIC:                             458.0
Df Residuals:                               16   BIC:                             464.6
Df Model:                                    5
Covariance Type:                     nonrobust
==========================================================================================================
                                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
Intercept                                -4.999e+05    5.3e+05     -0.944      0.359   -1.62e+06    6.23e+05
accesstoelectricity                       3234.5930   6190.213      0.523      0.608   -9888.071    1.64e+04
agriculturalland                          8709.2514   2225.300      3.914      0.001    3991.826    1.34e+04
co2emissionsliquid                           0.0077      0.511      0.015      0.988      -1.075       1.090
electricpower                               -3.0693     28.768     -0.107      0.916     -64.054      57.915
electricityproductionfromnaturalgassources  1470.0113    420.151      3.499      0.003     579.331    2360.692
==============================================================================
Omnibus:                        0.602   Durbin-Watson:                   1.376
Prob(Omnibus):                  0.740   Jarque-Bera (JB):                0.127
Skew:                          -0.184   Prob(JB):                        0.938
Kurtosis:                       3.053   Cond. No.                     2.51e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large. 2.51e+07. This might indicate that there are
```

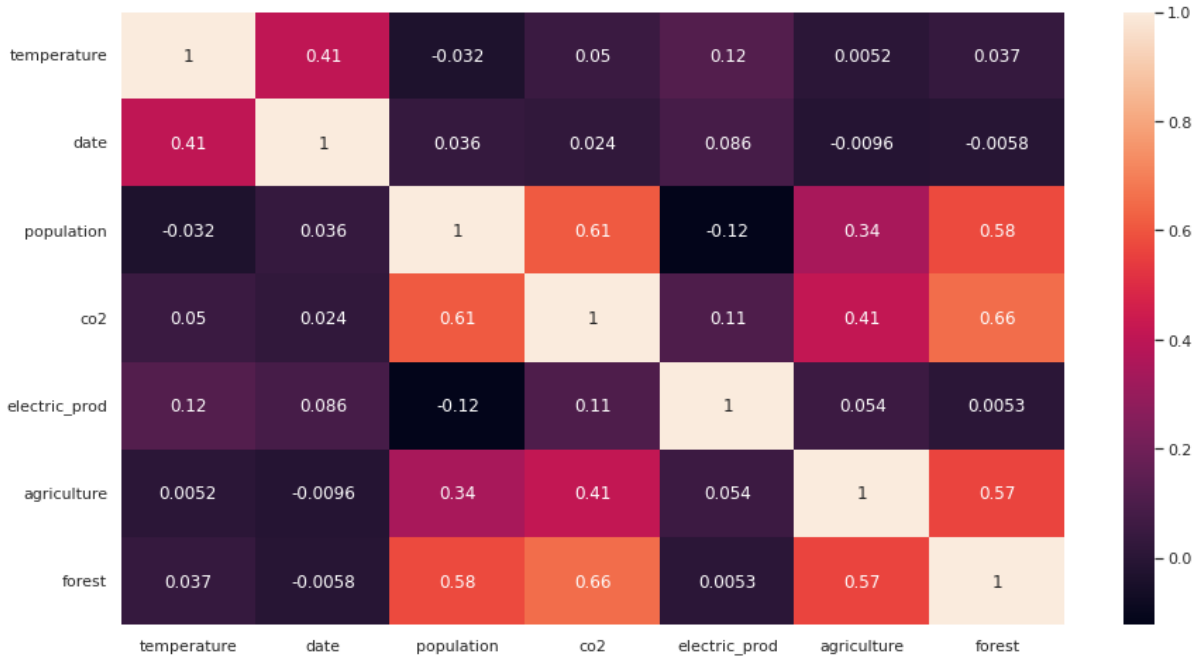
En esta sección se eligieron nuevas variables y el resultado de R^2 fue mejor obteniendo 0.97, lo que nos hace sospechar de un posible overfitting.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      temperaturechange   R-squared:                       1.000
Model:                             OLS   Adj. R-squared:                    nan
Method:                  Least Squares   F-statistic:                       nan
Date:                 Wed, 11 Jan 2023   Prob (F-statistic):                nan
Time:                         23:23:06   Log-Likelihood:                 761.58
No. Observations:                   22   AIC:                            -1479.
Df Residuals:                        0   BIC:                            -1455.
Df Model:                           21
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2330        inf          0        nan         nan         nan
date[T.1992]  -0.5330        inf         -0        nan         nan         nan
date[T.1993]  -0.1840        inf         -0        nan         nan         nan
date[T.1994]   0.2180        inf          0        nan         nan         nan
date[T.1995]  -0.0080        inf         -0        nan         nan         nan
date[T.1996]   0.2290        inf          0        nan         nan         nan
date[T.1997]   0.4770        inf          0        nan         nan         nan
date[T.1998]   0.0580        inf          0        nan         nan         nan
date[T.1999]  -0.1180        inf         -0        nan         nan         nan
date[T.2000]  -0.4520        inf         -0        nan         nan         nan
date[T.2001]   0.1620        inf          0        nan         nan         nan
date[T.2002]   0.0360        inf          0        nan         nan         nan
date[T.2003]   0.3930        inf          0        nan         nan         nan
date[T.2004]   0.1820        inf          0        nan         nan         nan
date[T.2005]  -0.0340        inf         -0        nan         nan         nan
date[T.2006]   0.2960        inf          0        nan         nan         nan
date[T.2007]  -0.4650        inf         -0        nan         nan         nan
date[T.2008]   0.2960        inf          0        nan         nan         nan
date[T.2009]   0.5510        inf          0        nan         nan         nan
date[T.2010]  -0.1550        inf         -0        nan         nan         nan
date[T.2011]   0.0910        inf          0        nan         nan         nan
date[T.2012]   0.5150        inf          0        nan         nan         nan
==============================================================================
Omnibus:                        0.033   Durbin-Watson:                   1.813
Prob(Omnibus):                  0.984   Jarque-Bera (JB):                0.086
Skew:                          -0.023   Prob(JB):                        0.958
Kurtosis:                       2.697   Cond. No.                         23.0
==============================================================================
```



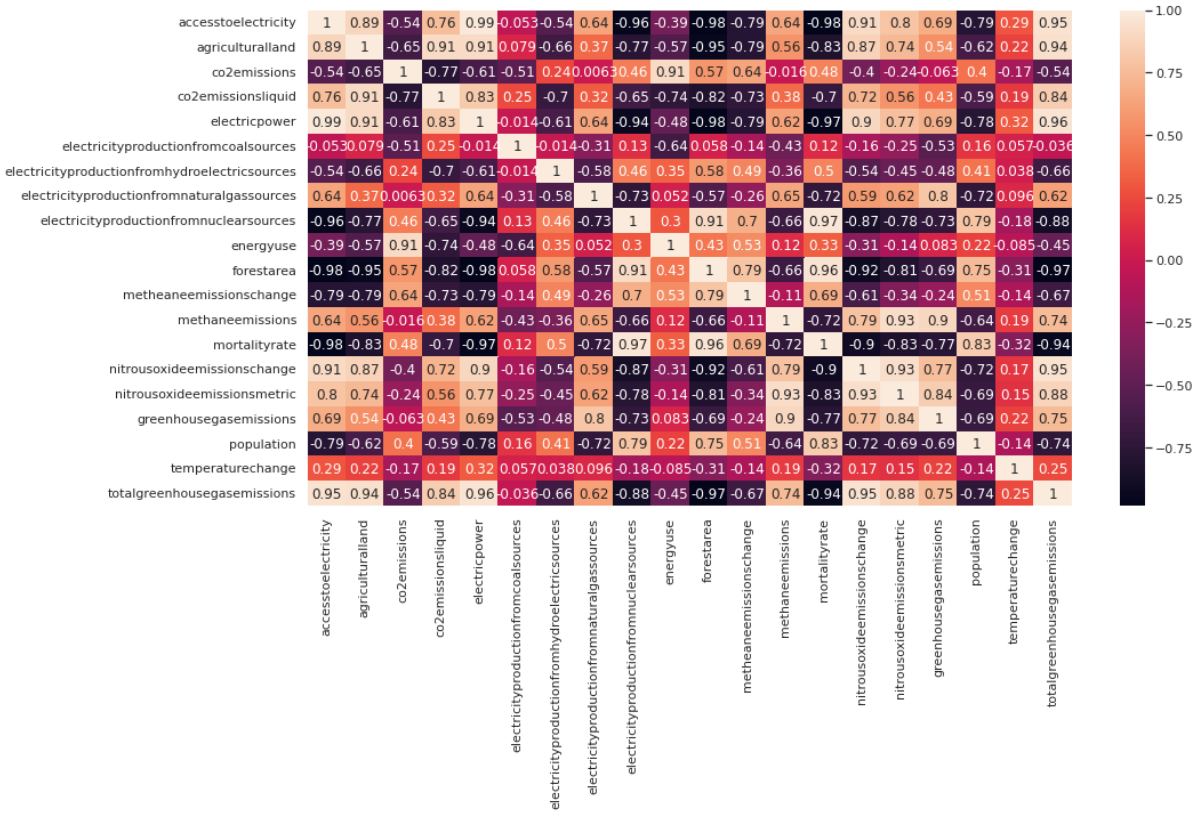
¿Qué actividad industrial/económica que haya emergido en el tiempo puede relacionarse con el cambio de temperatura?

Aparte del uso de energía eléctrica y el CO2 en el caso particular de Argentina destacan las actividades Agrícolas y la deforestación. La evolución de las ciudades ha hecho que las industrias alimentarias tengan que emplear métodos intensivos que pueden no ser sostenibles.



¿Cómo se comportan el resto de las variables?

Otra forma de ver el impacto de las variables entre sí es por medio de los mapas de calor. En este caso se pueden visualizar mejor las relaciones entre sí.



v) Preprocesamiento de Datos

Encoding.

No es necesario hacer una codificación de las variables ya que todas son continuas

vi) Splitting de dataset

En esta sección, hacemos la primera división con todas las variables para obtener los resultados de una regresión. Aún no hacemos la ingeniería de atributos ya que para nosotros esta regresión es preliminar y la ingeniería de atributos la vamos a usar con un algoritmo más robusto como XGBOOST y SVC.

Luego de los distintos analisis, nos quedamos con las siguientes variables:
-date
-accesstoelectricity
-agriculturalland
-co2emissions
-co2emissionsliquid
-electricpower
-electricityproductionfromcoalsources
-electricityproductionfromhydroelectricsources
-electricityproductionfromnaturalgassources
-electricityproductionfromnuclearsources
-energyuse
-forestarea
-metheaneemissionschange
-methaneemissions
-mortalityrate
-nitrousoxideemissionschange
-nitrousoxideemissionsmetric
- greenhousegasemissions
-population
-temperaturechange
-totalgreenhousegasemissions

Para comparar los valores de salida reales X_test con los valores predichos, convertimos en df:

	Actual	Predicted	Error	Error_porc
1	251,720.0	257,867.3	-6,147.3	-2.4
14	318,700.0	305,818.3	12,881.7	4.0
9	288,710.0	280,800.4	7,909.6	2.7
2	256,240.0	262,447.5	-6,207.5	-2.4
16	337,250.0	324,860.3	12,389.7	3.7

vii) Modelado de Machine Learning

Algoritmo Random Forest

Encontramos que el Random Forest tiene una función de regresión y es la que utilizamos como modelo.

Algoritmo SVR

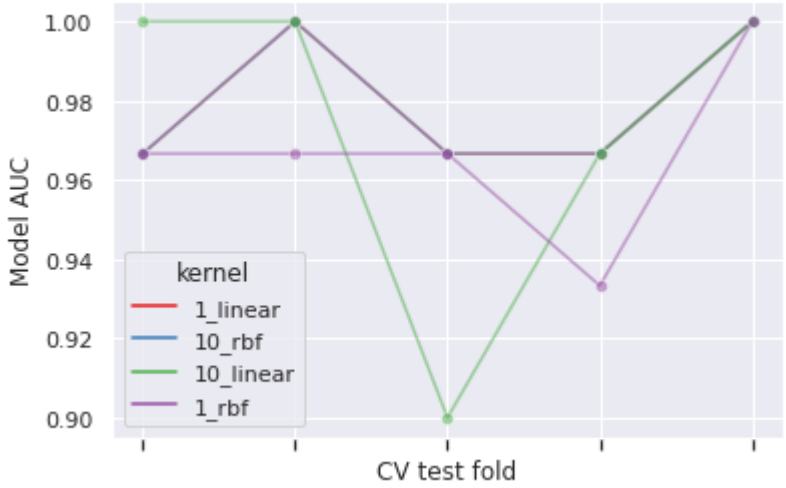
Como algoritmo para comparar queremos utilizar el SVR, Support Vector Regression

Usando el GridSearch pudimos obtener el best score

```
print('Best score: ', clf.best_score_)
```

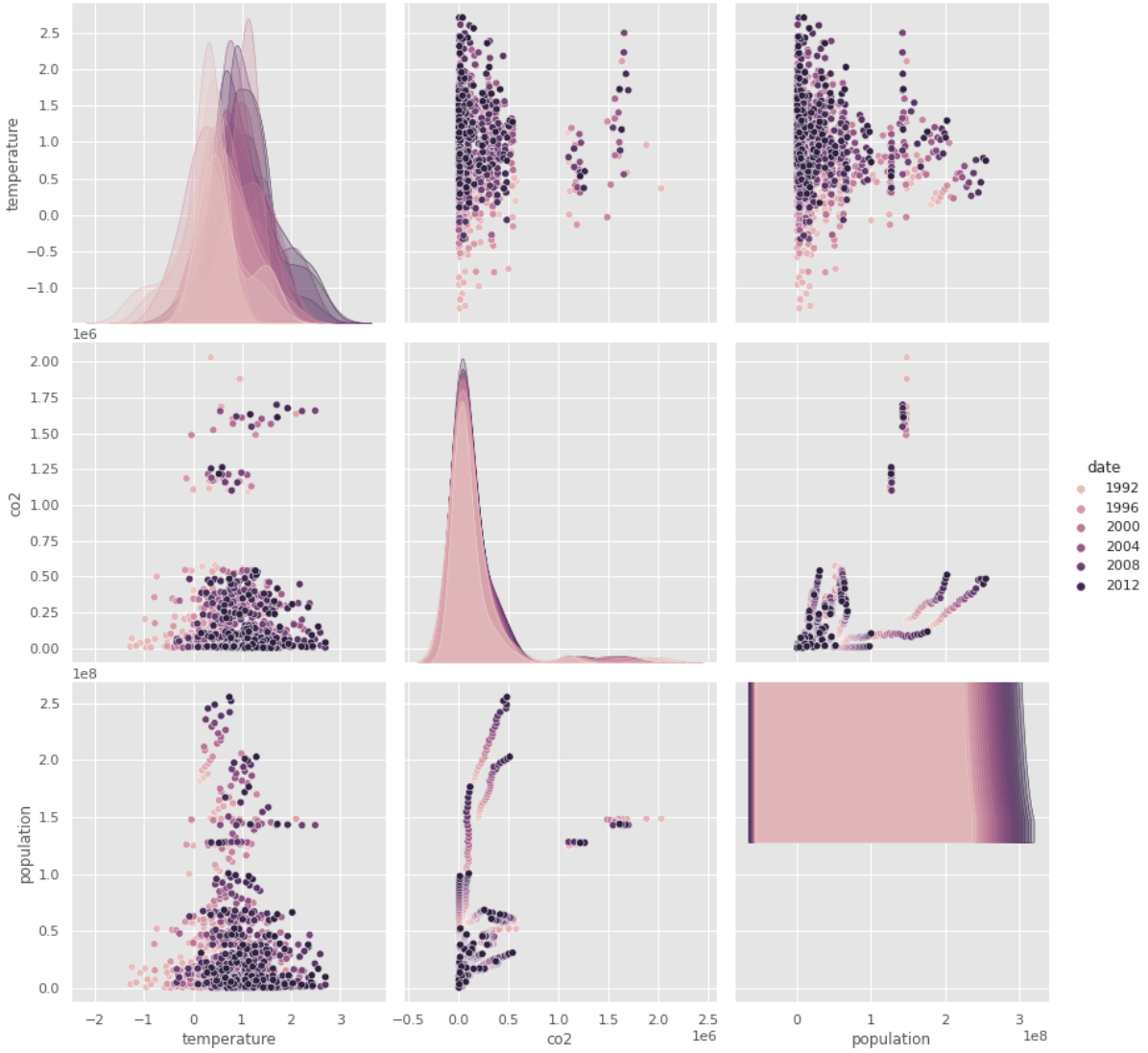
```
Best score:  0.9800000000000001
```

Este resultado nos hace dudar ya que el valor es muy cerano a 1 y esto nos puede dar indicios de que hay un overfitting.

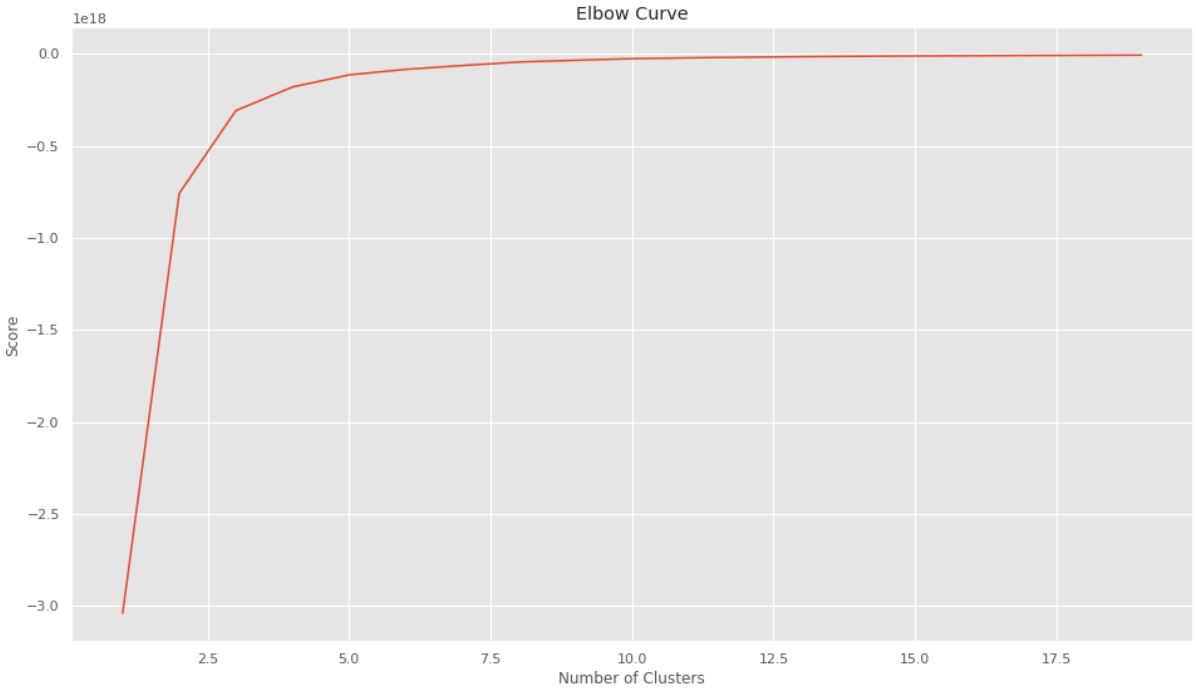


Algoritmo K-Means

Podemos observar que los grupos de datos (Clusteres) de 2012 para la temperatura y emisiones de co2 se agrupan hacia la derecha lo que denota un aumento en dicha variable:



Se elije el valor 5 de Clusters como un valor estable para aplicar este predictor.



viii) Explicación de Modelos

Visualización de Resultados de Modelo

Varianzas comunes

X1: Alrededor del 95% de la variabilidad de X1 se explica por los dos factores que seleccionamos.

X2: Alrededor del 90% de la variabilidad de X2 se explica por los dos factores que seleccionamos.

X3: Alrededor del 84% de la variabilidad de X3 se explica por los dos factores que seleccionamos.

X4: Alrededor del 87% de la variabilidad de X4 se explica por los dos factores que seleccionamos.

X5: Alrededor del 97% de la variabilidad de X5 se explica por los dos factores que seleccionamos.

X6: Alrededor del 66% de la variabilidad de X6 se explica por los dos factores que seleccionamos.

X7: Alrededor del 40% de la variabilidad de X7 se explica por los dos factores que seleccionamos.

X8: Alrededor del 67% de la variabilidad de X1 se explica por los dos factores que seleccionamos.

X9: Alrededor del 87% de la variabilidad de X2 se explica por los dos factores que seleccionamos.

X10: Alrededor del 86% de la variabilidad de X3 se explica por los dos factores que seleccionamos.

X11: Alrededor del 96% de la variabilidad de X4 se explica por los dos factores que seleccionamos.

X12: Alrededor del 70% de la variabilidad de X5 se explica por los dos factores que seleccionamos.

X13: Alrededor del 84% de la variabilidad de X6 se explica por los dos factores que seleccionamos.

X14: Alrededor del 95% de la variabilidad de X7 se explica por los dos factores que seleccionamos.

X15: Alrededor del 89% de la variabilidad de X2 se explica por los dos factores que seleccionamos.

X16: Alrededor del 83% de la variabilidad de X3 se explica por los dos factores que seleccionamos.

X17: Alrededor del 90% de la variabilidad de X4 se explica por los dos factores que seleccionamos.

X18: Alrededor del 69% de la variabilidad de X5 se explica por los dos factores que seleccionamos.

X19: Alrededor del 6% de la variabilidad de X6 se explica por los dos factores que seleccionamos.

```
fa.loadings_ # Matriz Factorial
```

```
array([[ 0.84260341, -0.4944043 ],
       [ 0.68324646, -0.65996803],
       [-0.12909534,  0.91268286],
       [ 0.51403189, -0.78041568],
       [ 0.82135449, -0.55257918],
       [-0.45418913, -0.67611338],
       [-0.523294  ,  0.36137948],
       [ 0.81955395,  0.08894578],
       [-0.86148705,  0.36734834],
       [ 0.02087687,  0.92952241],
       [-0.83034921,  0.5283838 ],
       [-0.42759896,  0.72382395],
       [ 0.90330408,  0.15732104],
       [-0.8944399 ,  0.38945635],
       [ 0.88210653, -0.34176733],
       [ 0.90706487, -0.11301518],
       [ 0.94150641,  0.13021436],
       [-0.79403327,  0.25326492],
       [ 0.20801191, -0.15310795]])
```

Segun la matriz factorial, las variables energyuse, forestarea, greenhousegasemissions y population definen el factor 1 (cargas altas en el factor 1, cargas relativamente pequeñas en el factor 2), mientras que las variables metheaneemissionschange, methaneemissions y nitrousoxideemissionschange definen el factor 2.

En nuestro análisis, el factor 1 representa los antecedentes de generación de electricidad y el factor 2 representa los antecedentes de generación de emisiones de co2 . Por lo tanto, podemos dar nombres relevantes para los dos factores de la siguiente manera.

Factor 1 → accesstoelectricity (variable relacionada con la generación de electricidad)

Factor 2 → co2emissions (variable relacionada con la generación de emisiones de co2).

Esto tiene sentido porque los gases de efecto invernadero está relacionado con el aumento del uso de electricidad y de la emisión de co2.

Varianzas comunes

-date: Alrededor del 95% de la variabilidad de date se explica por los dos factores que seleccionamos.

-accesstoelectricity: Alrededor del 90% de la variabilidad de date se explica por los dos factores que seleccionamos.

-agriculturalland: Alrededor del 84% de la variabilidad de agriculturalland se explica por los dos factores que seleccionamos.

-co2emissions: Alrededor del 87% de la variabilidad de co2emissions se explica por los dos factores que seleccionamos.

-co2emissionsliquid: Alrededor del 97% de la variabilidad de co2emissionsliquid se explica por los dos factores que seleccionamos.

-electricpower: Alrededor del 66% de la variabilidad de electricpower se explica por los dos factores que seleccionamos.

-electricityproductionfromcoalsources: Alrededor del 40% de la variabilidad de electricityproductionfromcoalsources se explica por los dos factores que seleccionamos.

-electricityproductionfromhydroelectricsources: Alrededor del 67% de la variabilidad de X1 se explica por los dos factores que seleccionamos.

-electricityproductionfromnaturalgassources: Alrededor del 87% de la variabilidad de accesstoelectricity se explica por los dos factores que seleccionamos.

-electricityproductionfromnuclearsources: Alrededor del 86% de la variabilidad de agriculturalland se explica por los dos factores que seleccionamos.

-energyuse: Alrededor del 96% de la variabilidad de co2emissions se explica por los dos factores que seleccionamos.

-forestarea: Alrededor del 70% de la variabilidad de co2emissionsliquid se explica por los dos factores que seleccionamos.

-metheaneemissionschange: Alrededor del 84% de la variabilidad de electricpower se explica por los dos factores que seleccionamos.

-methaneemissions: Alrededor del 95% de la variabilidad de electricityproductionfromcoalsources se explica por los dos factores que seleccionamos.

-nitrousoxideemissionschange: Alrededor del 89% de la variabilidad de accesstoelectricity se explica por los dos factores que seleccionamos.

-nitrousoxideemissionschange: Alrededor del 83% de la variabilidad de agriculturalland se explica por los dos factores que seleccionamos.

-greenhousegasemissions: Alrededor del 90% de la variabilidad de co2emissions se explica por los dos factores que seleccionamos.

-population: Alrededor del 69% de la variabilidad de co2emissionsliquid se explica por los dos factores que seleccionamos.

-temperaturechange: Alrededor del 6% de la variabilidad de electricpower se explica por los dos factores que seleccionamos.

ix) Insights y recomendaciones

El impacto del CO2 era predecible. Sin embargo, nos sorprendió el impacto de la actividad agrícola en el cambio de temperatura y el acceso a la electricidad ya que no precisamente es un factor relacionado directamente con la emisión de gases. Recomendaciones

Ya que conocemos el impacto que tienen las emisiones y uso de energía. Para palear el efecto de la temperatura, sería recomendable incentivar el uso de energías renovables ya que a medida que aumenta la población, se hace necesario producir electricidad.

En cuestión de la deforestación, se puede aprovechar el territorio para fomentar el crecimiento de la extensión forestal. También consideramos importante el control de la actividad agrícola para evitar efectos nocivos en el medio ambiente.

x) Fuentes

https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py

https://scikit-learn.org/stable/modules/model_evaluation.html?highlight=evaluation+measures

https://scikit-learn.org/stable/modules/model_evaluation.html?highlight=evaluation+measures

https://scikit-learn.org/stable/modules/model_evaluation.html?highlight=evaluation+measures

https://www.data-to-viz.com/

https://python-charts.com/es/?fbclid=IwAR2Um3KLimdSJ3BoufhVIqG69Z2Ooj5T_f1Awr8kuGQfgo3P7zEWUiZzYOw

https://scikit-learn.org/stable/modules/svm.html#regression

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

https://pythonbasics.org/split-train-test/

https://blog.bigml.com/2019/03/19/introduction-to-linear-regression/

https://scikit-learn.org/stable/modules/svm.html#regression

https://www.kaggle.com/code/mrisdal/exploring-survival-on-the-titanic/report

https://www.linkedin.com/posts/iamarifalam_kaggle-datasciences-projects-activity-6996086337376251904-cAAg/?utm_source=share&utm_medium=member_desktop

https://github.com/easystats/performance

https://lu.ma/STARTUP_10NOV

https://www.cienciadedatos.net/documentos/py39-forecasting-series-temporales-con-skforecast-xgboost-lightgbm-catboost.html

https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_stats.html#sphx-glr-auto-examples-model-selection-plot-grid-search-stats-py

https://scikit-learn.org/stable/modules/grid_search.html#grid-search

https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html#sphx-glr-auto-examples-svm-plot-svm-regression-py

https://www.geeksforgeeks.org/ordinary-least-squares-ols-using-statsmodels/

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

https://datascience.eu/es/matematica-y-estadistica/definicion-de-r-cuadrado/

https://stackoverflow.com/questions/50232599/interpreting-sklearns-gridsearchcv-best-score

https://www.datacamp.com/tutorial/principal-component-analysis-in-python

https://www.google.com/search?q=pca+python+example&oq=pca&aqs=chrome.3.69i57j0i67l3j46i199i465i512j0i512l5.4044j0j1&sourceid=chrome&ie=UTF-8

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

https://pub.towardsai.net/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa

https://www.jcchouinard.com/pca-with-python/

https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

https://www.geeksforgeeks.org/principal-component-analysis-with-python/

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://www.cienciadedatos.net/documentos/py19-pca-python.html

https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/

https://www.ciospain.es/big-data/que-son-los-datos-sinteticos-datos-generados-para-ayudar-a-tu-estrategia-de-ia

https://www.youtube.com/watch?v=8klqIM9UvAc

https://topbigdata.es/generacion-de-datos-sinteticos-mediante-una-red-generativa-de-adversarios-gan-con-pytorch-visual-studio-magazine/

https://www.datio.com/ai/datos-sinteticos-con-gans-i-por-que-datos-sinteticos/

https://sg.com.mx/dataday/synthdata-una-nueva-libreria-de-python-para-generacion-de-datasets-sinteticos/

https://pharos.sh/generacion-de-datos-sinteticos-con-numpy-y-scikit-learn/

https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

https://rua.ua.es/dspace/bitstream/10045/93556/1/Generacion_de_datos_sinteticos_para_segmentacion_Lopez_Avila_Placido_Antonio.pdf

https://www.linkedin.com/pulse/datos-sint%C3%A9ticos-y-python-hanna-alda/?originalSubdomain=es

https://www.youtube.com/watch?v=J1MxDETYoiU

https://www.kaggle.com/code/willkoehrsen/introduction-to-feature-selection

https://www.youtube.com/watch?v=mElVGah7Epg

https://www.datacamp.com/tutorial/principal-component-analysis-in-python

https://www.linkedin.com/posts/neel-shah-553105178_machine-learning-with-python-activity-7002267887109357568-ixC0/?utm_source=share&utm_medium=member_desktop

https://content.dataiku.com/ai-myths?utm_campaign=GLO+Content+AI+Myths+vs+Reality+Ebook+Aug+2021